nature ecology & evolution

ARTICLES
https://doi.org/10.1038/s41559-018-0478-6

Patterns of shared signatures of recent positive selection across human populations

Kelsey Elizabeth Johnson[1] and Benjamin F. Voight[2,3,4*]

Signatures of recent positive selection often overlap across human populations, but the question of how often these overlaps represent a single ancestral event remains unresolved. If a single selective event spread across many populations, the same sweeping haplotype should appear in each population and the selective pressure could be common across populations and environments. Identifying such shared selective events could identify genomic loci and human traits important in recent history across the globe. In addition, genomic annotations that recently became available could help attach these signatures to a potential gene and molecular phenotype selected across populations. Here, we present a catalogue of selective sweeps in humans, and identify those that overlap and share a sweeping haplotype. We connect these sweep overlaps with potential biological mechanisms at several loci, including potential new sites of adaptive introgression, the glycophorin locus associated with malarial resistance and the alcohol dehydrogenase cluster associated with alcohol dependency.

Positive selection is the process whereby a genetic variant rapidly increases in frequency in a population due to the fitness advantage of one allele over the other. Recent positive selection has been a driving force in human evolutinn, and studies of loci targeted by positive selection have uncovered potential adaptive phenotypes in recent human evolutionary history (for example, refs [1–3]). One observation that has emerged from scans for positively selected loci is that these signatures often overlap across multiple populations, localized to discrete locations in the genome[4–7]. Sequencing data available from diverse human populations provide an opportunity to characterize the frequency that overlapping signatures share a common, ancestral event—and potentially a common selective pressure. Identifying shared selective events would be of fundamental interest, highlighting loci and traits important in recent history across the globe.

Functional annotations across the human genome can also potentially connect variants targeted by selection with candidate genes and an associated mechanism. For example, the influx of expression quantitative trait loci (eQTL) across many tissue types[8], and inferred regions of ancient hominin introgression[9], now provide a richer foundation to investigate the potential biological targets under selection. While identifying the causal variant at a site of positive selection is notoriously difficult, if single nucleotide polymorphisms (SNPs) on a selected haplotype are associated with changes in expression of a nearby gene, this information could help attach the signature to a potential gene and molecular phenotype.

Here, we focus on the detection of genomic signatures compatible with selection on a newly introduced mutation that has not yet reached fixation (that is, a hard, ongoing sweep) to explore their distribution across populations and spanning the genome. We performed a scan for positive selection using the integrated haplotype score (iHS) on 20 populations from four continental groups from phase 3 of the 1000 Genomes Project (1KG)[10]. We found that 88% of sweep events overlapped across multiple populations; correlating with population relatedness and geographic proximity; 59% of overlaps were shared (that is, a similar sweeping haplotype was present) across populations; and 29% of overlaps were shared across continents. We connect these multi-population sweep overlaps with potential mechanisms at (1) the glycophorin cluster (*GYPA*, *GYPB* and *GYPE*), where we observe sweeps across all four continental groups in a region associated with malarial resistance; (2) sweeps across African populations at the X chromosome gene *DGKK*, implicated in hypospadias in males; (3) a sweep shared in European populations tagged by a coding variant in the gene *MTHFR*, which is associated with homocysteine levels and a multitude of additional traits; (4) two putative regions of adaptive introgression from Neandertals; and (5) the alcohol dehydrogenase (*ADH*) cluster, where a sweep in Africa is associated with alcohol dependence in African Americans.

Results and discussion

Our broad objective was to apply the iHS to identify and characterize the distribution of selective sweeps across human populations and the genome. Towards this end, we first evaluated the distribution of iHS when applied to whole-genome sequencing data.

A correction to iHS adjusting for local, low recombination rates. The iHS was conceptualized for population data ascertained for common genetic variation[11], and may not be fully calibrated for sequencing data that include rare variation. To examine the score in more detail, we initially applied the iHS to genome sequencing data obtained from 1KG (Methods). We observed an excess of SNPs tagging strong iHS signals at lower derived allele frequencies (Supplementary Fig. 1a) in a frequency range where iHS is not expected to have substantial power[11] (note that a negative iHS indicates extended haplotype homozygosity on the derived relative to the ancestral allele). We observed a negative correlation between the number of populations in an overlap and the local recombination

[1]Genetics and Gene Regulation Program, Cell and Molecular Biology Graduate Group, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [4]Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.
*e-mail: bvoight@pennmedicine.upenn.edu

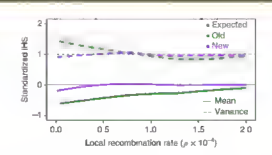


Fig. 1 | Normalizing the iHS by local recombination rate. The mean (solid line) and variance (dashed lined) of the iHS as a function of local recombination rate. iHS values normalized by derived allele frequency are shown in green; normalization by both derived allele frequency and local recombination rate is shown in purple. The grey lines represent the mean (=0) and variance (=1) for comparison.

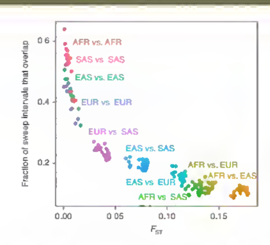


Fig. 2 | Closely related populations have sweep overlaps more frequently. For each population pair, the fraction of sweep intervals that overlap is plotted against pairwise estimated F_{ST}. Each population pair (dots) is coloured by its continental groupings (for example, EUR vs. SAS = one European population versus one South Asian population). AFR, African; EAS, East Asian; EUR, European, SAS, South Asian.

rate in any population (for example, Pearson's correlation = −0.24, $P<8.5\times10^{-10}$ for the CHB population; see Supplementary Table 2 for population codes). Simulations showed that normalizing the iHS only by derived allele frequency misses additional homozygosity found at low-frequency variants in 'cooler' recombination rate regions, resulting in scores that are not entirely cross-comparable (Supplementary Fig. 2). Normalizing the iHS both by derived allele frequency and local recombination rate using a binning approach (Methods) resulted in better calibrated standardized iHS (Fig. 1), which substantially reduced the excess of high-scoring iHS values at low frequencies (Supplementary Fig. 1b). This approach also eliminated correlation between low recombination rates and sweep overlaps (for example, Pearson's correlation = −0.04, $P=0.35$ for the CHB population after correction). In the following, we utilize this normalization scheme for the iHS genome wide, treating autosomes separately from the X chromosomes (Methods and Supplementary Table 1).

A catalogue of signals of recent positive selection across human populations. We measured the iHS normalized separately for the autosomes and X chromosome across 26 populations from the 1KG project (Methods and Supplementary Table 2). For each population's iHS scan, we identified putative sweep intervals that segregated an unusual aggregation of extreme iHS values (Methods and Supplementary Table 3). Consistent with previous reports[11], the number of sweep intervals per population correlated with its effective population size (Supplementary Table 4). We defined the tag SNP for each interval as the highest scoring variant by the absolute value of the iHS, as we expect the tag SNPs to be in strong linkage disequilibrium with the putative causal, selected variant of the sweep. Our sweep intervals recovered 11 of the 12 top signatures reported in the original iHS study[11], and 14 of the top 22 signatures reported elsewhere[12]. We observed more extreme iHS using whole genome sequencing data compared with array-based genotype data. For example, previously in the CEU population only 6 of 256 signatures (2.3%) had an $|iHS|>5$ for their most extreme score[11]; in our scan, 92 of 597 signatures (15%) had an $|iHS|>5$. This observation is consistent with a more rapid decay of homozygosity on background haplotypes due to improved ascertainment of low-frequency and rare variants across the genome. We next sought to characterize the frequency with which putative sweep intervals overlap the same genomic region across multiple populations. We excluded recently admixed populations (ASW, ACB, MXL, PEL, PUR, CLM) as events observed in those groups may reflect selection in ancestral populations pre-dating admixture (Methods). Consistent with previous reports[4–6], we noted that related populations (as measured by the population differentiation statistic F_{ST}) more often overlap in their putative sweeps intervals relative to more distantly related pairs (regression coefficient $\beta=-2.06$, s.e. = 0.52, $P=2.5\times10^{-61}$; Fig. 2).

Overlapping sweep intervals across human populations occur in genomic hotspots. Through characterizing sweep overlaps, we observed cases where many sweeps appeared to cluster in specific genomic locations. The most striking clustering of overlapping sweep intervals occurred on chromosome 17 (Fig. 3a), with 23 overlapping events in total, of which 14 span continental groups. While previous reports have investigated how often sweeps overlap across the globe, to our knowledge, the extent to which putative sweep intervals are organized across the genome has not been quantified. To model this phenomenon, we fit the observed rate of overlapping sweep events in 10 Mb windows with individual or mixtures of Poisson distributions (Methods). We first modelled the count of genes in each window. We did not expect a single Poisson process to be the best fit to the gene count in each window, as genes are not uniformly distributed across the genome. Indeed, we found that a mixture model with five components best fit the frequency at which genes occur in the genome (Methods). Next, we found that the counts of sweeps in an individual population in genomic windows were best modelled by a single rate (Fig. 3b). This result was surprising, as we expected that sweeps targeting genetic variation in genes or nearby regulatory regions would follow a non-uniform rate like genes themselves. When we looked at the rate of sweep overlaps, we found that sweeps overlapping across populations were best fit by a mixture of Poisson distributions with three different rates ($P=2.7\times10^{-7}$, χ^2 test versus two-component mixture, Fig. 3c and Methods). These results suggest that some regions of the genome have a higher rate of overlaps than others, but this observation was not obviously explained by the number of genes in a window (Pearson's correlation = 0.07, $P=0.22$). We also tested the relationships of local recombination rate and estimates of background selection[13] with the number of sweep overlaps in a window. We hypothesized that these features might improve a model of regions of the genome under selection rather beyond just the locations of genes, and could better explain the observed hotspots of sweep overlaps. While both features were correlated with overlap count, local recombination rate was most significant (~5% variance explained) and local background selection did not explain additional variation. These data indicate that putative sweep intervals overlapping multiple human populations may aggregate in discrete 'hotspots' of activity, which are not entirely explained by

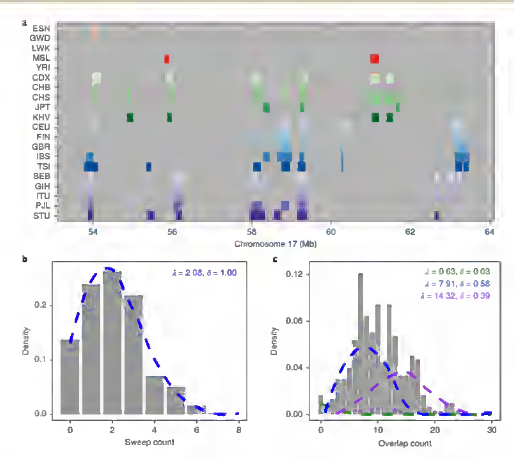


Fig. 3 | Overlapping sweeps tend to cluster in the genome. a, An example of a 10 Mb window on chromosome 17 with multiple overlaps across many populations. See Supplementary Table 2 for population codes. **b**, The distribution of sweep interval counts in 10 Mb windows across the genome for a single population (LWK). The histogram plots the observed counts, and the blue dashed line is the best-fit Poisson distribution. **c**, The distribution of sweep overlaps across two or more populations in 10 Mb windows across the genome. The histogram plots the observed counts, and the dashed lines represent the results of Poisson mixture modelling. The best-fit model was the three-component model shown here. λ, Poisson distribution rate, δ, mixture proportion.

the presence of genes, background selection or local recombination rates. These hotspots could be targeting specific genic regions, or additional genomic features not assessed here.

Complex patterns of sweep sharing across populations and continents. We next sought to identify selective sweeps that are potentially shared across populations, that is, where the putative sweeping haplotype is similar across populations. Sharing could occur in several ways, including a common ancestral event occurring before population divergence that persisted to the present day, or via gene flow of advantageous alleles between populations. To characterize haplotype similarity across populations at our genomic intervals tagged by unusual iHS, we utilized the program fastPHASE[18]. Using a hidden Markov model, fastPHASE models the observed distribution of haplotypes as mosaics of K ancestral haplotypes, allowing us to map a sweep tag SNP to an ancestral haplotype jointly across multiple populations at once without arbitrarily choosing a physical span to build a tree of haplotypes or otherwise measure relatedness (Methods).

Overall, out of 1,803 intervals shared across populations, 521 (29%) were shared across continents, frequently between Europe and South Asia, consistent with observed lower genetic differentiation relative to other continental comparisons (Supplementary Table 5). Indeed, consistent with our previous analysis using all intervals, F_{ST} predicted the fraction of sweep overlaps that were shared between a pair of populations ($\beta = -2.80$, s.e. $= 0.31$, $P = 9.1 \times 10^{-88}$; Supplementary Fig. 3). Though more closely related populations have a higher fraction of shared sweeps, they also have more total unshared sweeps. This relationship could be due to shared selective pressures in nearby populations, false negatives in our sharing analysis or a combination of both.

To determine whether the observed extent of sweep sharing was unusual, we applied our fastPHASE haplotype labelling procedure to matched random sites across the genome. For all within-continent population pairs, and all but 4 of 75 Eurasian between-continent pairs, the degree of sweep sharing was higher than the background rate (Fig. 4 and Supplementary Table 6), suggesting that the sweep sharing we observed was not driven purely by haplotype similarities across closely related populations. The number of populations in a shared sweep was inversely correlated with the length of the shared haplotype (Spearman's rho $= -0.23$, $P < 2.2 \times 10^{-16}$, Supplementary Fig. 4), which is compatible with more widely shared sweeping haplotypes being broken down by recombination over time. There was a borderline significant correlation between the nucleotide diversity of a sweeping haplotype and the number of populations sharing a sweep (Spearman's rho $= 0.049$, $P = 0.041$, Supplementary Fig. 5), though this weak correlation was not significant when we looked at diversity between populations within continents.

Though the majority of between-continent shared sweeps were found across non-African populations, we did observe examples of shared sweeps between African and non-African populations. In total, 9.4% of observed sweep overlaps between African and

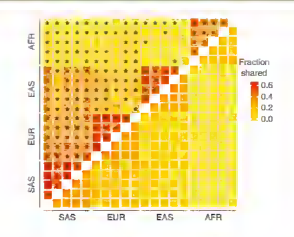


Fig. 4 | Enrichment of shared sweeps across population pairs. Squares below the diagonal represent the null fraction of overlaps shared across population pairs, from randomly placed overlaps across the genome. Squares above the diagonal represent the observed fraction of sweep overlaps shared for each population pairs. Squares are marked with an asterisk if the observed fraction shared was significantly higher than the null distribution. Populations are arranged alphabetically within continental groups by population code, top to bottom, right to left.

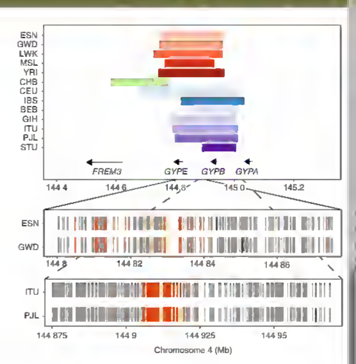


Fig. 5 | Signatures of positive selection at the *GYP* locus on chromosome 4. We observed signatures of positive selection in 13 populations at the *GYP* locus, including at least one population from each studied continental group. The top panel shows the sweep intervals of populations with sweeps at this locus, and the positions of the *GYP* genes. The bottom panels show the sweeping haplotypes for two African (ESN, GWD) and two South Asian (ITU, PJL) populations' within-continent shared sweeps. The grey tick marks in each population's row indicate the presence of a derived allele on the sweeping haplotype most common in that population, with a black tick indicating the position of each population's SNP with the most extreme iHS value. Also shown in orange are the significant eQTL for *GYPE* (light orange) or both *GYPB* and *GYPE* (dark orange) in linkage disequilibrium with these populations' shared haplotype (linkage disequilibrium measure $D'=1$). The eQTL for *GYPB* and *GYPE* are from whole blood.

non-African population pairs were called as shared (491 total), compared with 4.0% of control overlaps (99% confidence interval: 3.7–4.4%). For example, on chromosome 1 at ~47 Mb, a sweeping haplotype shared across African and European populations fell in a cluster of cytochrome P450 genes (Supplementary Fig. 6 and Supplementary Note).

Shared and overlapping sweeps in a region implicated in malaria resistance. With a catalogue of shared and overlapping selective sweeps in hand, we next aimed to identify specific regions of sweep sharing that connected the interval to a gene, pathway or phenotype when considered alongside additional genomic annotations. With a sweep overlap across thirteen populations from all four continental groups, the glycophorin gene cluster (*GYPA*, *GYPB*, *GYPE*) came to our attention for its repeated targeting by positive selection and its previous implication in malaria resistance (Fig. 5). This genomic region has been noted as a target of positive selection in humans[15–18], and as a target of ancient balancing selection shared between humans and chimpanzees[19]. In the IBS and South Asian populations, the sweep appeared to be on a shared haplotype, while the African, CHB and CEU populations each had unique sweeping haplotypes (Fig. 5). This complex locus contains a segmental duplication, making mapping and phasing of short-read data difficult. However, we observed residual unusual iHS in the surrounding region, and in an iHS scan of only those variants passing the 1KG 'strict' mask. We identified multiple signatures of positive selection on distinct haplotypes in all four continental groups (Supplementary Table 5), with some linked to one or more potentially causal variant(s), either coding or structural (Supplementary Note). The frequency and diversity of apparent adaptive pressures at this locus underscores the role of selection on host–pathogen interactions over recent and longer evolutionary timescales in modern humans, and the potential importance of this locus in particular in that process.

Intersection of signatures of positive selection with the GWAS catalogue. Previous work has indicated an enrichment of extreme iHS at genome-wide association studies (GWAS) signals for autoimmune diseases[20], and we hypothesized that this or other traits might be enriched for GWAS signatures linked to our signatures of positive selection. We tested for enrichment of linked GWAS SNPs overall and for specific traits (autoimmune disorders, height and schizophrenia) by generating random SNP sets from the HapMap3 variant set (Methods), and identifying the number of sweeps in each population linked to a GWAS variant compared with the random SNP sets. We saw no clear, compelling evidence of enrichment for GWAS SNPs overall or any of the traits tested. In total, 186 sweep tag SNPs from all 20 populations (out of 11,655; 1.6%) were in strong linkage disequilibrium with at least one genome-wide significant GWAS SNP ($r^2 \geq 0.9$, Supplementary Table 7), compared with a mean of 193 in our random SNP sets (95% confidence interval: 168–220). However, this intersection did identify candidates for a potential phenotype under selection at some loci. In one example, a sweep overlap across all five African populations falls at the gene *DGKK* on the X chromosome (Supplementary Note and Supplementary Fig. 7). Variants in this gene have been associated in Europeans with hypospadias[21,22], a prevalent birth defect of ectopic positioning of the opening of the urethra in males. A second example occurred at the *MTHFR* gene on chromosome 1, where a non-synonymous variant (A222V, rs1801133) has been extensively studied for its association with homocysteine levels[23,24]. A sweep overlap at this locus with three European populations (CEU, GBR, IBS) and JPT was called as shared across all four populations (Supplementary Note and Supplementary Fig. 8).

Evidence for adaptive introgression from Neandertals in non-African populations. Examples of positive selection on introgressed genetic variation have shown that positive selection acted on genetic variation from ancient hominins at some loci. While some of these examples are confined to a single population (for example, *EPAS1* in Tibetans[25]), most are common across multiple populations[26,27], and thus we hypothesized that a subset of our shared sweeps could be examples of adaptive introgression. We identified 141 candidate sweeps in partial to strong linkage disequilibrium with inferred introgressed haplotypes[9] ($r^2 \geq 0.6$, Methods and Supplementary Table 8), including previously described adaptive targets such as the *HYAL2* locus in East Asians[28] and *OAS1* in Europeans[29]. We did not observe an overall enrichment of these introgressed haplotypes in our iHS intervals ($P = 0.59$, χ^2 test, Methods), suggesting that introgression alone was not predictive of an unusual iHS signature. Of these 141 loci, we illustrate two candidate sweeps shared across multiple populations (Supplementary Note): (1) a shared sweep between Europeans and South Asians on chromosome 3 near *CT64*, a non-coding RNA primarily expressed in the testes (Supplementary Fig. 9); and (2) a sweep at ~41 Mb on chromosome 1, where all five South Asian populations have evidence of an introgressed haplotype at low to moderate frequency (18–30%) (Supplementary Fig. 10).

Overlapping and shared sweeps enriched in the ethanol oxidation pathway. We next sought to explore possible biological pathways targeted by shared selective events. As a large fraction of causal variants under positive selection are potentially non-coding[30–32], we hypothesized that regulatory variation in the form of eQTL could indicate a potential causal, functional variant and/or gene target. We identified genes with *cis* eQTL from all tissue types in the GTEx V6p dataset that were linked with shared sweeps ($r^2 \geq 0.9$) and tested for over-representation of biological pathways in this set of genes using ConsensusPathDB[33]. Excluding the human leukocyte antigen (HLA) genes (Methods), the most significant pathway was ethanol oxidation ($P = 2 \times 10^{-5}$, false discovery rate q value = 0.047), with seven of ten genes included in our shared sweeps gene set. This pathway includes the *ADH* gene cluster, which contains a previously described East Asian selective event targeting rs1229984-T (ref. [34]), a derived non-synonymous variant in *ADH1B* associated with increased *ADH1B* enzyme activity[35] and decreased risk of alcohol dependence in East Asians[36]. A recent report also found evidence for an independent selective event for the variant rs1229984 in Europeans[37]. Within this region, we observed the East Asian sweep, and independent sweeping haplotypes in the YRI and ESN populations (Fig. 6).

As GWAS have identified genetic variation in the *ADH* locus associated with alcohol dependence[36,38], we tested whether these associations were linked to the sweeping haplotype. In the YRI sweep interval spanning *ADH1B*, the derived allele of the leading iHS SNP (rs12639833-T, iHS = −5.133) was significantly associated with decreased risk for alcohol dependence in African Americans[38]. This alcohol dependence GWAS in African Americans identified independent associations at a non-synonymous variant in *ADH1B* (rs2066702) and a synonymous variant in *ADH1C* (rs2241894)[38]. The sweep tag in YRI is in perfect linkage disequilibrium with the *ADH1C* lead variant rs2241894 ($r^2 = 1$), lies in an intron of *ADH1C*, and is a significant eQTL for *ADH1C* (esophagus mucosa) and *ADH4* (esophagus muscularis, skeletal muscle). Several other SNPs in strong linkage disequilibrium with rs12639833 have extreme negative iHS, are eQTLs for increased *ADH1C* and *ADH4* expression (including in the liver)[8,39], and were significantly associated with decreased risk for alcohol dependence in African Americans (Supplementary Table 9). Alcohol dehydrogenases oxidize ethanol to acetaldehyde, a process that is thought to occur primarily in the liver[40]. These data suggest a similar mechanism is at play in individuals

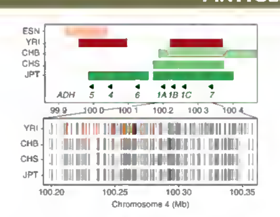


Fig. 6 | Signatures of positive selection at the *ADH* locus on chromosome 4. The top panel shows the sweep intervals of populations with sweeps at this locus, and the positions of the seven *ADH* cluster genes. The bottom panel shows the sweeping haplotypes of the four populations with sweeps in this region, with grey tick marks indicating the derived alleles present on the most common sweeping haplotype in that population. The black tick marks indicate the position of the SNP with the most extreme iHS in each population. For the YRI population, the positions of significant *ADH4* and *ADH1C* eQTL in subcutaneous adipose tissue (light orange), GWAS SNPs from ref. [38] (dark orange), and SNPs that are eQTL for both genes and are GWAS SNPs (red) in linkage disequilibrium with YRI's tag SNP ($r^2 > 0.9$) are shown.

of West African ancestry as in East Asians, where the selected allele increases *ADH* enzyme activity[35], resulting in an adverse physical response to alcohol consumption[41], and reduced risk for alcohol dependence[36]. Taken collectively, these patterns suggest that (1) alcohol oxidation pathways broadly have been subject to recent positive selection in humans, (2) that genes in this pathway have been repeatedly targeted, with multiple events segregating at these sites, (3) the selective pressure appears to operate across the major continental groups included in this study, and (4) sweeping haplotypes at the *ADH* locus tag functional variation associated with protection against alcohol dependence.

In summary, we identified overlapping and shared signatures of positive selection across human populations, using a modified version of the iHS statistic. We observed more extreme iHS in sequencing data compared with SNP array genotype data, which could be a consequence of more rapid decay of homozygosity on unselected haplotypes due to the presence of rare variants. We found that closely related populations are more likely to share sweeping haplotype signatures, though we identified examples of sharing across genetically distant populations. These loci immediately raise questions of how these examples arose, whether by gene flow after divergence or a common ancestral event. Though only a small amount of gene flow between African and non-African populations is thought to have occurred since their divergence, the introduction of an adaptively advantageous allele at very low frequency could lead to the signature we observed. But in considering the collection of putative shared sweeps we highlighted here, it seems apparent that each locus is unique, segregating patterns of genetic variation suggestive of a range of compatible (and potential quite complex) population models that could help explain these data. Future work to infer the potential scenarios leading to shared sweeps will likely require modelling of individual regions to elucidate the evolutionary history of specific events. We also found that the rate of sweep overlaps is not uniform across the genome, but in some locations overlaps cluster together, contributing to the complexity of the underlying sweep signatures in those regions. These features made identifying the tag SNP for a sweep and calling sharing between sweep overlaps difficult in these regions. That said, we hope that our catalogue of

unusually long haplotypes shared across human populations will help to elucidate genes—and ultimately phenotypes—that are still evolving across the wide range of environments human have experienced in recent history.

Methods

iHS scan. We downloaded phased genotype files for phase 3 of the 1000 Genomes Project from the 1KG FTP server (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). These data were converted to Beagle-formatted files, and filtered to include only biallelic single nucleotide variants (excluded indels) with a minor allele frequency (MAF) greater than 1%. A fine-scale recombination map was downloaded from the 1KG FTP server (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130507_omni_recombination_rates/), and scaled to units of ρ ($=4N_e r_c$, where r_c is the recombination rate per generation) for each population. Effective population size (N_e) was estimated for each population by calculating nucleotide diversity (π) in a sliding window (100 kb) across the genome, and estimating N_e from the median values of π ($N_e=\pi/4\mu$, where μ is the mutation rate per generation). Ancestral alleles were identified using the human–chimp–macaque alignment from Ensembl (accessed from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/). SNPs were filtered for only those where the ancestral allele was supported by both the chimp and macaque alignments.

Unstandardized iHS were calculated using WHAMM (v0.14c), using a modified version of iHS calculation code (v1.3) that increased speed of the calculation, and initially standardized by derived allele frequency as described in the original iHS paper[11], with 50 allele frequency bins. In the final standardization, we binned autosomal SNPs into 500 bins (50 allele frequency bins × 10 local recombination rate bins), or 150 bins for chromosome X (50 allele frequency bins × 3 local recombination rate bins). These standardization files are available in Supplementary Table 1.

Regions of the genome putatively undergoing recent hard sweeps—what we refer to in the main text as iHS intervals—were identified by counting the number of SNPs with $|iHS|>2$ in 100 kb windows (windows incrementing by one SNP, that is, overlapping windows). We took the union of the top 1% of windows, by the total number or by fraction of SNPs with $|iHS|>2$ in the window, as our intervals. We performed this interval calling separately for each of the 20 populations included in this study. The SNP we used to label (that is, tag) each sweep interval was identified as the SNP with the most extreme iHS, and the sweep frequency as the tag SNP-derived allele frequency if the iHS was less than zero, and ancestral allele frequency if iHS was greater than zero. We limited our analyses of individual sweep loci to those with a tag SNP of MAF > 15%, to focus on signatures unlikely to have extreme iHS due to very low frequency.

Neutral simulations. We performed neutral simulations of a population with a CHB population-like demographic history using the forward simulation software SLiM[42] (v. 2.4.2). We simulated chromosomes with the following demographic model[43]: an ancestral population size of 13,000 diploids with a mutation rate of 1×10^{-7} was burned in for 130,000 generations, then a bottleneck reduced the population size to 2,500 diploids at 32,490 generations, followed by exponential growth at a rate of 1.002082, and finally 216 diploids were sampled at generation 34,500. We simulated 85, 3 Mb regions at a recombination rate of $r_c=1\times10^{-8}$ or 1×10^{-12}, and calculated unstandardized iHS for all variants. Only variants in the central 200 kb were used for the comparison in Supplementary Fig. 2, to maximize the number of low-frequency (<5% MAF) variants that reached the end of their haplotype within the simulated chromosomes (for a total of 159,487 variants at $r_c=1\times10^{-8}$ and 151,710 variants at $r_c=1\times10^{-12}$). The mean of unstandardized iHS for variants in an allele frequency bin were compared between the difference recombination rates with the Mann–Whitney test (wilcox.test), and the variances were compared with the F test (var.test) in R[44].

Sweep overlaps. To identify sweep overlaps, we compared the iHS intervals for each population and identified regions of the genome where two or more populations had a sweep interval. We calculated the fraction of sweep overlaps for each population pair as the mean of the fraction of sweep intervals in one population that overlap with a sweep interval in the second population (that is, (fraction in population A + fraction in population B) / 2). We estimated F_{ST} for each pair of populations across all variants ($n=2,627,240$) in the 1000 Genomes VCF files on chromosome 2 using the Weir and Cockerham estimator implemented in VCFtools (v. 0.1.12b)[45]. We performed linear regression on (fraction of overlaps ~ F_{ST}) for each population pair, and estimated the standard error of the slope using a block jackknife for unequal group size[46], deleting one population pair group (for example, EUR versus EUR, or AFR versus EUR) for each resample.

Rates across the genome. To assess the rate of sweep intervals across the genome, we subdivided the genome into 10 Mb non-overlapping windows ($n=297$ in total) and counted the number of sweep intervals for each individual population, and the number of overlaps across two or more populations, in each window. To ensure the sweep intervals called for each population were independent, we merged adjacent sweep intervals into one interval if their tag SNPs were in modest linkage disequilibrium or greater ($r^2>0.4$). If a sweep interval or overlap spanned two windows, we counted it once in the window with more than half of its physical distance. We used all Ensemble HG19 gene annotations (from http://genome.ucsc.edu/cgi-bin/hgTables), merged into non-overlapping intervals with BEDTools v2.19.1 (ref. [47]). We used the local recombination rates described above, and background selection estimates (B statistic) were downloaded from http://www.phrap.org/software_dir/mcvicker_dir/bkgd.tar.gz13. We performed forward regression with the medians of B and local recombination rate in our windows as potential predictors of overlap count in each window, along with the gene count in the window. The final model included only median local rho ($R^2=0.048$, $P=5.2\times10^{-5}$), as median B explained no additional variation. For Poisson mixture modelling of the overlap rate, we fit mixtures of independent Poisson distributions to the data by minimizing the negative log likelihood with the nonlinear minimizer function (nlm) in R[44]. We compared mixture models by calculating the Bayesian information criterion and performing a likelihood ratio test.

Identifying sweeping haplotypes with fastPHASE. For each sweep overlap, we identified the physical region spanning all tag SNPs, and an additional 5 kb to either side. We ran fastPHASE on this region, using the -u option to identify each 1KG population as a subpopulation, -B to indicate known haplotypes and -Pzp to output cluster probabilities for each individual at each SNP. We tested a range of values of K (number of haplotype clusters) and T (number of random expectation maximization algorithm starts) on a subset of sweep overlaps, and found broadly similar results across the range (Supplementary Fig. 11). We used $K=10$ clusters and $T=10$ for all overlaps in the final analysis. From the output cluster probabilities, we identified the sequence of haplotype clusters for each SNP position in each individual as the most likely haplotype cluster at each SNP. We then identified the haplotype cluster sequences of all chromosomes carrying the selected tag allele, and the most common of those to be the reference sweeping haplotype sequence.

To identify a pair of populations as 'shared', we required an identical reference haplotype sequence to span the selected tag allele in both populations. To form shared clusters, we grouped together all populations that were called as shared with at least one other population. To calculate the null rate of haplotype sharing across population pairs, we selected random regions of the genome of the same size (within 10 kb), distance to the nearest gene (within 50 kb), and local recombination rate (within an order of magnitude of $\rho=4N_e r_c$) as our observed sweep overlap regions. For each sweep overlap, we identified 10 matched windows, for a total of 30,450 regions across the genome (ranging from 153 to 2,588 random overlaps per population pair). We identified tag SNPs for each population in the random regions matching the distance from the other populations' tag SNPs and derived allele frequency (within 5%) of the observed overlap. We then ran fastPHASE on the randomly selected regions and performed the shared haplotype-calling procedure as for observed overlap windows described above. To compare the observed fraction of overlaps called as shared to the null haplotype sharing for each pair of populations, we performed 1,000 bootstraps by sampling with replacement the number of observed overlaps from the null. Population pairs where the shared sweep fraction of observed overlaps was higher than the shared fraction of random overlaps for all 1,000 samples are marked with an asterisk in Fig. 4. We performed linear regression on (fraction of shared overlaps ~F_{ST}) for each population pair, and estimated the standard error of the slope using a block jackknife for unequal group size[46], deleting one population pair group (for example EUR versus EUR, or AFR versus EUR) for each resample.

We measured the length of the shared sweeping haplotype in each population as the maximum length of the reference shared haplotype, identified as above, in that population. We calculated nucleotide diversity in each population as the average number of pairwise differences between chromosomes carrying the tag allele in the region of the shared haplotype, divided by the length of the shared haplotype in that population. We tested for correlations between the number of populations sharing a sweep and the mean shared haplotype length and mean nucleotide diversity using Spearman's rho in R (cor.test). We also tested, in each population individually, the correlation between nucleotide diversity of the shared haplotype in that population and the number of populations in the shared sweep.

eQTL linked to sweep haplotypes. To connect shared sweeps to potential causal genes, we utilized the GTEx v6p eQTL dataset[8] downloaded from the GTEx portal (http://www.gtexportal.org/). For each population's tag SNPs, we identified linkage disequilibrium proxies ($r^2\geq0.9$, calculated in the same population) within 1 Mb of the sweep interval, and intersected these SNPs with all significant GTEx eQTL from all tissue types. eQTL in the GTEx V6p dataset were identified using a cohort of mostly white individuals (84.3%), with a smaller fraction of African Americans (13.7%). For sweep overlaps that were called as shared, we identified a shared SNP set as the intersection of linkage disequilibrium proxy sets for all populations in a shared group. We created a gene list of all genes with eQTL from any tissue that intersected with shared SNP sets, excluding HLA genes. We chose to exclude HLA genes, owing to its genomic complexity and its enrichment for signatures of recent positive selection. To test for enrichment of this gene set with biological pathways,

we used over-representation analysis of all pathway databases in ConsensusPathDB (http://cpdb.molgen.mpg.de/)[33] with the background set of all genes.

Intersection of sweeps with Neandertal haplotypes, Neandertal PheWAS and GWAS SNPs. We downloaded the Neandertal haplotype calls reported in ref. [48] from http://akeylab.gs.washington.edu/vernot_et_al_2016_release_data/. This dataset contains inferred introgressed Neandertal haplotypes on the autosomes of the non-African individuals from 1KG phase 3. To calculate linkage disequilibrium between introgressed haplotypes and sweep tag SNPs, we pooled overlapping haplotypes across individuals and created a genotype of 0 or 1 based on presence or absence of the overlapping introgressed haplotype in each individual. We then calculated linkage disequilibrium between this presence or absence genotype and the tag SNPs within 1 Mb of the introgressed haplotype separately for each population. We considered haplotypes with $r^2 > 0.6$ with sweep tag SNPs as candidates for adaptive introgression. To examine potential enrichment of introgressed haplotypes in linkage disequilibrium with sweep tag SNPs, we compared the fraction of introgressed haplotypes in linkage disequilibrium with sweep tag SNPs to the fraction of distance and frequency matched SNPs in linkage disequilibrium with sweep tag SNPs ($r^2 > 0.6$) using a χ^2 test. We downloaded the Neandertal phenome-wide association studies (PheWAS) data at https://phewascatalog.org/neanderthal[49], and intersected all reported associations with variants in strong linkage disequilibrium ($r^2 \geq 0.9$) with each sweep tag SNP in each population.

We downloaded the GWAS catalogue from https://www.ebi.ac.uk/gwas on 12 October 2016. We identified all genome-wide significant associations ($P < 5 \times 10^{-8}$) in strong linkage disequilibrium ($r^2 \geq 0.9$) with each sweep tag SNP in each population. To test for enrichment of GWAS variants generally and of specific phenotype classes, we performed permutation tests with 10,000 random SNP sets from the HapMap3 variant set (from ftp://ftp.ncbi.nlm.nih.gov/hapmap/phase_3/) matched for allele frequency and distance to gene with the GWAS variants of interest. We then compared the empirical distribution of intersection of these matched SNP sets with the sweep tag SNPs and proxies to the number of observed GWAS intersections. To control for potentially linked GWAS variants, we simply counted the number of sweeps in each population that intersected a GWAS or control set variant.

Indels, structural variants and annotations. Indels were not included in our iHS scan, but could be the causal variant on a sweeping haplotype. To identify candidates for causal indels, we calculated linkage disequilibrium with sweep tag SNPs for all indels in the 1000 Genomes phase 3 VCF files within 1 Mb of the sweep interval in each population. To identify potential functional coding variants among indels and SNPs on sweeping haplotypes, we used ANNOVAR to annotate coding variation[50]. In the glycophorin region, we tested for linkage disequilibrium between sweep tag SNPs and structural variant calls from phase 3 of the 1KG (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/; 24 February 2015 release).

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. WHAMM v0.14c and iHS calculator v1.3 are available at http://coruscant.itmat.upenn.edu/whamm/dlnotes.html. Custom code used to process and analyse output from the iHS scan and fastPHASE is available at https://github.com/kelsj/sharedSweeps.

Data availability. The standardized iHS values for all 26 populations in 1KG phase 3 are available at http://coruscant.itmat.upenn.edu/data/JohnsonEA_iHSscores.tar.gz.

Received: 28 June 2017; Accepted: 11 January 2018;
Published online: 19 February 2018

References

1. Tishkoff, S. A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* **39**, 31–40 (2007).
2. Kamberov, Y. G. et al. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* **152**, 691–702 (2013).
3. Fumagalli, M. et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* **349**, 1343–1347 (2015).
4. Pickrell, J. K. et al. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* **19**, 826–837 (2009).
5. Coop, G. et al. The role of geography in human adaptation. *PLoS Genet.* **5**, e1000500 (2009).
6. Metspalu, M. et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am. J. Hum. Genet.* **89**, 731–744 (2011).
7. Liu, X. et al. Detecting and characterizing genomic signatures of positive selection in global populations. *Am. J. Hum. Genet.* **92**, 866–881 (2013).
8. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
9. Vernot, B. & Akey, J. M. Resurrecting surviving neandertal lineages from modern human genomes. *Science* **343**, 1017–1021 (2014).
10. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
11. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
12. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
13. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
14. Scheet, P. & Stephens, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* **78**, 629–644 (2006).
15. Baum, J., Ward, R. H. & Conway, D. J. Natural selection on the erythrocyte surface. *Mol. Biol. Evol.* **19**, 223–229 (2002).
16. Wang, H. Y., Tang, H., Shen, C. K. J. & Wu, C. I. Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites. *Mol. Biol. Evol.* **20**, 1795–1804 (2003).
17. Ko, W. Y. et al. Effects of natural selection and gene conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. *Am. J. Hum. Genet.* **88**, 741–754 (2011).
18. Leffler, E. M. et al. Resistance to malaria through structural variation of red blood cell invasion receptors. *Science* **356**, eaam6393 (2017).
19. Leffler, E. M. et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* **339**, 1578–1582 (2013).
20. Raj, T. et al. Common risk alleles for inflammatory diseases are targets of recent positive selection. *Am. J. Hum. Genet.* **92**, 517–529 (2013).
21. van der Zanden, L. F. M. et al. Common variants in DGKK are strongly associated with risk of hypospadias. *Nat. Genet.* **43**, 48–50 (2011).
22. Geller, F. et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. *Nat. Genet.* **46**, 957–963 (2014).
23. Paré, G. et al. Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma Homocysteine in a healthy population a genome-wide evaluation of 13 974 participants in the women's genome health study. *Circ. Cardiovasc. Genet* **2**, 142–150 (2009).
24. van Meurs, J. B. J. et al. Common genetic loci influencing plasma homocysteine concentrations and their effect on risk of coronary artery disease. *Am. J. Clin. Nutr.* **98**, 668–676 (2013).
25. Huerta-Sánchez, E. et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* **512**, 194–197 (2014).
26. Racimo, F., Sankararaman, S., Nielsen, R. & Huerta-Sánchez, E. Evidence for archaic adaptive introgression in humans. *Nat. Rev. Genet.* **16**, 359–371 (2015).
27. Racimo, F., Marnetto, D. & Huerta-Sánchez, E. Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* **34**, 296–317 (2017).
28. Ding, Q., Hu, Y., Xu, S., Wang, J. & Jin, L. Neanderthal introgression at chromosome 3p21.31 was under positive natural selection in east asians. *Mol. Biol. Evol.* **31**, 683–695 (2014).
29. Sams, A. J. et al. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. *Genome Biol.* **17**, 246 (2016).
30. Grossman, S. R. et al. Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–713 (2013).
31. Enard, D., Messer, P. W. & Petrov, D. A. Genome-wide signals of positive selection in human evolution. *Genome Res.* **24**, 885–895 (2014).
32. Yu, F. et al. Population genomic analysis of 962 whole genome sequences of humans reveals natural selection in non-coding regions. *PLoS One* **10**, e0121644 (2015).
33. Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Res.* **37**, D623–D628 (2009).
34. Han, Y. et al. Evidence of positive selection on a class I ADH locus. *Am. J. Hum. Genet.* **80**, 441–456 (2007).
35. Edenberg, H. J. The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res. Health* **30**, 5–13 (2007).
36. Li, D., Zhao, H. & Gelernter, J. Strong association of the alcohol dehydrogenase 1B gene (ADH1B) with alcohol dependence and alcohol-induced medical diseases. *Biol. Psychiatry* **70**, 504–512 (2011).
37. Galinsky, K. J. et al. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet* **98**, 456–472 (2016).

38. Gelernter, J. et al. Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry* **19**, 41–49 (2014).
39. Pashos, E. E. et al. Large, diverse population cohorts of hiPSCs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci. *Cell Stem Cell* **20**, 558–570.e10 (2017).
40. Lee, S. L., Chau, G. Y., Yao, C. T., Wu, C. W. & Yin, S. J. Functional assessment of human alcohol dehydrogenase family in ethanol metabolism: significance of first-pass metabolism. *Alcohol. Clin. Exp. Res.* **30**, 1132–1142 (2006).
41. Matsuo, K. et al. Alcohol dehydrogenase 2 His47Arg polymorphism influences drinking habit independently of aldehyde dehydrogenase 2 Glu487Lys polymorphism: analysis of 2,299 Japanese subjects. *Cancer Epidemiol. Biomark. Prev.* **15**, 1009–1013 (2006).
42. Haller, B. C. & Messer, P. W. SLiM 2: flexible, interactive forward genetic simulations. *Mol. Biol. Evol.* **34**, 230–240 (2017).
43. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2016).
44. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2016).
45. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
46. Busing, F. M. T. A., Meijer, E. & Leeden van der, R. Delete-m jackknife for unequal m. *Stat. Comput.* **9**, 3–8 (1999).
47. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
48. Vernot, B. et al. Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals. *Science* **352**, 235–239 (2016).
49. Simonti, C. N. et al. The phenotypic legacy of admixture between modern humans and Neandertals. *Science* **351**, 737–741 (2016).
50. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

Acknowledgements

We thank C. Brown and E. Leffler for helpful comments that improved the quality of the manuscript. This work was supported by grants from the National Institutes of Health (NIDDK R01DK101478 to B.F.V., T32GM008216 to K.E.J.) and a fellowship from the Alfred P. Sloan Foundation (BR2012-087 to B.F.V.).

Author contributions

K.E.J. and B.F.V. planned the study. K.E.J. assembled input data and performed the experiments. K.E.J. and B.F.V. interpreted the data and wrote the paper.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-018-0478-6.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to B.F.V.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.